House Price Case Study Subjective Questions and Answers

Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

The optimal value of alpha for Ridge regression is 4.

The optimal value of alpha for Lasso regression is 0.0001.

- In Regularization alpha determines how severe the penalty is.
- When we double alpha in Ridge Regression the slope gets asymptotically close to zero i.e. the co-efficient values will be reduced such that the values are close to zero.
- In contrast, when the value of alpha is doubled in Lasso Regression the slope will be shrieked to zero i.e. the redundant features values will be reduced to zero. There by providing feature selection.
- From the model the above can be more clearly explained as below:
- Out of 300 Features:
    - At the optimal value:
        - Ridge picked 287 variables and eliminated the other 11 variables.
        - Lasso picked 208 variables and eliminated the other 90 variables.
    - After the alpha was doubled:
        - Ridge picked 287 variables and eliminated the other 11 variables.
        - Lasso picked 183 variables and eliminated the other 115 variables.

Clearly, it is visible that when the penalty increases, Lasso Regression reduces the effect of redundant variables by shrinking their values to zero.

The most important predictor variables after the change is implemented are :

Ridge Regression:

1. OverallCond_9
2. Neighborhood_Crawfor
3. OverallQual_9

4. OverallCond_8
5. Neighborhood_StoneBr

Lasso Regression:

1) SaleType_Oth
2) FullBath_3
3) BsmtFullBath_2
4) Exterior1st_BrkFace
5) RoofMatl_Tar&Grv

Question 2

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer:

There are two scenarios which helps us to decide between Ridge and Lasso regression in real World use cases:

- Ridge Regression is used when we don't want to remove any features i.e. all the features are very important.
- Also when the number of parameters are more compared to the number of samples. Example when we want to build model for all 10,000 parameters with only 500(or fewer samples)
- Lasso Regression is useful when we want do feature selection i.e. when we want to determine the most important feature that determines the predictor variable and shrinkage.

In this scenario, we want to determine the most important features that decide the price of the house. Not all the features might lead to proper results there might be redundant features too. The parameters are less compared to the number of samples. We need regularization as well as feature selection in short. In such cases Lasso Regression best fits the needs and determines only those features that affect the price of the house. LASSO regression or L1-norm penalty, sets some of the model coefficients to exactly zero instead of just shrinking them. Effectively, this does the 'automatic feature selection' i.e. let's you automatically ignore the unimportant features even if you start with a highly complex model to fit the data. As per Lasso regression the top features that determine the high price of house are:

1. OverallQual_10
2. OverallQual_9
3. OverallCond_9
4. BsmtFullBath_2
5. FullBath_3

Question 3

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:

The five most important predictor variables are:

1) SaleType_Oth : Sale type with tag as others
2) Neighborhood_StoneBr : Physical location of the house is in Stone Brook.
3) Neighborhood_Crawfor : Physical location of the house is in Stone Brook
4) MSZoning_FV : Zoning Classification is Floating Village Residential
5) RoofMatl_Tar&Grv : Roof material Gravel & Tar

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer:

According to Occam 's Razor boundary the model should not be too complex or too simple. Advantages of the model being simple will be robust (i.e. any changes in the train data the model doesn't change significantly) and generalised (i.e. performs better on the unseen data). Both of these are achieved using Lasso Regression. This is, in fact, one of the key advantages of LASSO regression or L1-norm penalty, that it sets some of the model coefficients to exactly zero instead of just shrinking them. Effectively, this does the 'automatic feature selection' for you i.e. let's you automatically ignore the unimportant features even if you start with a highly complex model to fit the data.

This is clearly visible in the House Price Case Study, Initially as per the optimum value of alpha the Lasso regression identifies few features as important predictors, but when we remove those features and again run the Lasso Regression for the model, the model again selects another few features as the important predictors. There by making the model generalised and robust.

More accurate model leads to over fitting. A flexible one will learn from data a lot. So usually avoiding accurate model and preferring a robust one is better.